Safety Theatre and the Suppression of Agency

How Risk Discourse Becomes a Control Mechanism

Author: Anthony Paterson
Affiliation: Fractal Media Infrastructure (FMI)
Contact: https://github.com/instance001
Version: 0.1 (2026-02-10)

Abstract

In recent decades, the concept of safety has expanded beyond the management of material risk to encompass behavioral, linguistic, and cognitive regulation. This paper examines the emergence of safety theatre as a structural phenomenon increasingly oriented not toward harm reduction, but toward constraining discretionary judgment outside sanctioned procedures. Contemporary safety frameworks often prioritize procedural compliance, moral signaling, and liability insulation over situational reasoning and competence development.

While framed as protective, these systems can reallocate responsibility upward and erode decision-making capacity downward, producing environments in which actors learn to defer rather than to model and reason. The paper identifies mechanisms through which safety theatre reduces opportunities for agentic decision-making, describes characteristic fragilities of compliance-dependent institutions, and argues that disciplined agency is a precondition of resilient safety.

Positioning (scope)

This paper does not argue that contemporary safety frameworks are universally misguided, nor that regulation is inherently incompatible with agency. Rather, it examines a specific structural drift observable across multiple domains, wherein safety mechanisms increasingly function to constrain decision-making authority rather than to cultivate judgment under uncertainty. The analysis focuses on systemic outcomes rather than intent, and on recurring institutional patterns rather than isolated failures.

Keywords

Safety theatre; agency; risk governance; institutional design; competence; decision-making

1. Introduction: When Safety Stops Being About Harm

Historically, safety referred to the mitigation of concrete risks: physical injury, system failure, or environmental hazard. Such frameworks assumed uncertainty, contextual judgment, and tradeoffs. In contrast, contemporary safety discourse increasingly focuses on regulating behavior rather than managing risk, with emphasis placed on procedure, language, and adherence to prescribed norms.

This shift marks a structural transformation. Safety is no longer defined primarily by outcomes, but by compliance with sanctioned processes. As a result, the central question moves from “What action best reduces harm in this context?” to “Who is authorized to decide, and under what constraints?”

The claim advanced here is not that safety mechanisms have no value, but that some safety mechanisms increasingly function as authority-allocation systems: they distribute (and deny) discretionary judgment. In doing so, they can reduce the conditions under which competence is learned and exercised.

2. Redefining Safety Theatre

Classically, “theatre” in safety or security refers to measures that create the appearance of protection without meaningfully reducing risk. The modern variant differs in an important respect: it is often effective—effective at behavioral constraint rather than safety itself.

We can define modern safety theatre as:

> Ritualized safeguards that substitute procedural compliance and moral signaling for contextual judgment, while presenting this substitution as ethical care.

Under this definition, safety theatre is not merely bureaucratic excess or incompetence. It is a system-level pattern in which procedure becomes a proxy for responsibility, and deviation—regardless of competence or intent—is treated as a hazard in itself. The conceptual lineage is close to “security theatre” discussions in security studies and public discourse (e.g., Schneier 2003), but the object here is broader: the governance of discretion.

3. Mechanism I: Responsibility Reallocation

A core mechanism of modern safety theatre is the upward reallocation of responsibility. Decision-making authority is centralized into policies, frameworks, and abstract oversight bodies, while individuals are positioned as execution layers.

In this arrangement:

- Outcomes are subordinated to adherence.
- Accountability is transferred from the actor to the procedure.
- “Following policy” becomes both justification and defense.

This reallocation reframes discretion as a risk factor. Individual judgment is no longer treated as a safety asset but as a liability that introduces variance the system is designed to suppress. Responsibility is not eliminated; it is displaced—away from those closest to the context and toward

structures that are legible and governable, but often too abstract to respond to local reality (Scott 1998).

In bounded-rational environments, this displacement can appear attractive: procedure is a tractable substitute for decision-making under uncertainty (Simon 1955). The cost is that local judgment becomes harder to practice and easier to punish.

4. Mechanism II: Moral Signaling as Legitimizing Infrastructure

Safety theatre increasingly relies on moral language to legitimate constraint. Concepts such as care, harm prevention, and ethical responsibility can function not primarily to guide action, but to stabilize institutional legitimacy in the face of uncertainty and blame.

This produces a distinctive pattern:

- Ethical language substitutes for empirical justification.
- Compliance is framed as virtue.
- Objection is reframed as recklessness or harm endorsement.

Under these conditions, disagreement is not evaluated on its merits but on its alignment with sanctioned moral framing. Discourse shifts from problem-solving to boundary enforcement. The result is not necessarily more safety, but more predictable adherence—often aligned with liability minimization and reputational protection (Power 2004).

5. Mechanism III: Competence Suppression Loops

The most consequential effect of safety theatre is the formation of self-reinforcing competence suppression loops.

The loop is straightforward:

1. Discretion is restricted to reduce perceived risk.
2. Reduced discretion limits opportunities to practice judgment.
3. Diminished practice leads to reduced competence.
4. Reduced competence is cited as justification for further restriction.

Over time, systems that follow this pattern become dependent on low-agency participants to function smoothly. Competence becomes anomalous rather than expected, and discretion appears dangerous precisely because it is no longer well-developed.

6. Failure Mode Inversion

Compliance-oriented systems often appear stable. They perform adequately under routine conditions, where variability is low and scenarios match predefined templates. However, this apparent stability masks a critical vulnerability.

When novel, ambiguous, or high-stakes situations arise, these systems can fail disproportionately. The suppression of discretionary judgment leaves participants less able to adapt, escalate, or improvise. Responsibility bottlenecks at higher levels, while local actors lack both authority and practice to respond.

This inversion—routine success paired with exceptional fragility—helps explain why audit- and procedure-heavy environments can look “safe” until they do not. Related patterns are documented in institutional accidents and organizational failure literatures (e.g., Vaughan 1996; Perrow 1984).

7. Agency as a Safety Primitive

Contrary to prevailing assumptions, agency is not the enemy of safety. Untrained, unbounded agency may increase risk, but disciplined, contextual agency is a prerequisite for resilient systems.

Agency enables:

- Context-sensitive decision-making
- Error detection and correction
- Adaptation under uncertainty

Safety emerges not from the absence of agency, but from its cultivation under appropriate constraints. Systems that treat agency as a liability may reduce visible risk in the short term, but they can accumulate latent fragility.

8. Cross-Domain Implications

Although this analysis is domain-general, its implications are observable across multiple sectors:

- Education: emphasis on rule-following over understanding.
- Healthcare: protocol adherence displacing clinical judgment.
- Bureaucracy: process compliance prioritized over service outcomes.
- Organizations: risk aversion replacing initiative.
- Technology governance: restriction favored over capability development.

In each case, safety theatre produces similar tradeoffs: apparent order at the cost of adaptive capacity.

9. Conclusion: Safety Without Agency Is a Dead System

Safety frameworks that constrain discretionary judgment may appear humane, ethical, and orderly. Yet by externalizing responsibility and eroding competence, they can create systems that are increasingly brittle. When conditions shift—as they inevitably do—the cost of suppressed agency becomes visible all at once.

Genuine safety does not arise from obedience alone. It depends on people who are permitted—and prepared—to think, judge, and act within constraints. Systems that forget this may avoid risk temporarily, but they do so by trading away resilience itself.

Appendix A. Scope and non-claims

To prevent misreadings, the argument does not require any of the following claims:

- That all safety mechanisms are theatre.
- That regulators or institutions are acting in bad faith.
- That safety is less important than autonomy.
- That constraints are never appropriate.

The argument does require a narrower claim: that some institutional safety mechanisms can be well-explained as tools for allocating and constraining discretion, and that this shift can generate predictable deskilling and brittleness.

Appendix B. Relation to risk and governance literature (brief)

The paper’s mechanisms overlap with several established lines of work:

- Legibility and local knowledge: centralization trades local judgment for

governable simplifications (Scott 1998).

- Risk and audit cultures: institutions can optimize for demonstrable compliance

and defensible process over adaptive performance (Power 2004).

- Bounded rationality: procedures are attractive substitutes for decision-making

under uncertainty, but can become rigid proxies for responsibility (Simon 1955).

- Organizational failure: routine normalization and complexity-driven accidents

expose the limits of compliance-only safety (Vaughan 1996; Perrow 1984).

References

- Perrow, C. (1984). Normal Accidents: Living with High-Risk Technologies. Princeton University Press.
- Power, M. (2004). *The Risk Management of Everything: Rethinking the Politics of Uncertainty*. Demos.
- Schneier, B. (2003). *Beyond Fear: Thinking Sensibly About Security in an Uncertain World*. Copernicus Books.
- Scott, J. C. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press.
- Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1), 99–118.
- Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press.